doi: 10.12194/j.ntu.20210507001
引文格式: 文万志, 姜文轩, 葛威, 等. 一种基于深度学习的实体消歧技术[J]. 南通大学学报(自然科学版), 2021, 20(4): 23-30.

一种基于深度学习的实体消歧技术

文万志, 姜文轩, 葛　威, 朱　恺, 李喜凯, 吴雪斐
(南通大学　信息科学技术学院, 江苏　南通　226019)

摘要: 传统的命名实体消歧技术通常依靠丰富的上下文语境和外部实体知识库,而很多新兴实体缺乏知识库且包含实体的文本长度较短,这些局限性使得传统算法不能够充分利用上下文的语义信息。另外,由于受有效样本数量的限制,算法最终应用的场景十分有限。基于上述问题,提出一种基于深度学习的结合 BERT(bidirectional encoder representation from transformers)模型和长短期记忆神经网络的实体消歧方法。该方法主要包含以下几个部分:1)设计了一种基于 BERT 模型的词向量,通过较少的数据样本仍然可以获取较多的信息;2)为了让长短期记忆神经网络保留较多的有用信息和验证短文本以适用该方法,对句子样本进行切分;3)结合微软公司提出的 NNI(neural network intelligence)技术,高效地获取较优的神经网络超参数。通过与其他不同类型的词向量和神经网络技术进行比较,验证了使用文中基于深度学习的实体消歧技术在 F-Measure 值评测指标上效果更好。

关键词: 深度学习; 自然语言处理; 实体消歧; 长短期记忆; 神经网络
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673-2340(2021)04-0023-08

An Entity Disambiguation Method Based on Deep Learning

WEN Wanzhi, JIANG Wenxuan, GE Wei, ZHU Kai, LI Xikai, WU Xuefei
(School of Information Science and Technology, Nantong University, Nantong 226019, China)

Abstract: The traditional named entity disambiguation technology usually relies on rich context and knowledge of external entities. However, many emerging entities lack knowledge bases and the text containing entities is short. These limitations make traditional algorithms unable to make full use of contextual semantic information. At the same time, due to the limitation of the number of effective samples, the final application scenarios of the algorithm are very limited. Based on the above defects, this paper proposes a deep learning-based entity disambiguation method combining bidirectional encoder representation from transformers (BERT) model and long short-term memory neural network. The main work are the following parts: 1) A word vector based on the BERT model is designed to obtain more information through fewer data samples. 2) In order to allow the long short-term memory neural networks to retain useful information and verify that the short text applies to the method of this article, this method segments the sentence samples. 3) This article uses the neural network intelligence (NNI) technology proposed by Microsoft, which makes it possible to quickly and efficiently obtain the optimal neural network hyperparameter. This study compares other different types of word vectors and neural network technology, confirming that the F-Measure value of the entity disambiguation technology based on deep learning used in this paper is higher.

Key words: deep learning; natural language processing; entity disambiguation; long short-term memory; neural network

收稿日期: 2021-05-07
基金项目: 国家自然科学基金项目(61602267);工业信息化部重点实验室开放基金项目(NJ2018014)
第一作者简介: 文万志(1982—), 男, 副教授, 博士, 主要研究方向为机器学习、软件测试与调试、程序切片技术。E-mail:wenwanzhi@126.com

南通大学学报(自然科学版)

命名实体消歧在自然语言处理领域发挥着十分重要的作用,其目的是解决文本中实体歧义问题。一般而言,命名实体在文本信息传输过程中发挥着关键作用,但命名实体通常以简称的方式存在,这可能导致多个实体指向一个相同的实体名称,也就是所谓的实体歧义。实体消歧的任务就是将文本中的实体正确地链接到实体语义中。实体消歧作为自然语言处理领域的基础性研究,对后续的语言处理任务十分重要,相关任务包括:智能问答[1]、信息降噪[2]、人工智能翻译[3]等。

近些年,实体消歧技术在不断进步,其实用性、适用性和稳定性不断提高。目前,实体消歧技术包含了机器学习和通过维基百科构建的语料库[4]实现的大数据技术。邵发等[5]针对开放文本中中文实体关系抽取的一词多义问题,提出一种基于实体消歧的中文实体关系抽取方法。通过在知网中挖掘关系实体构建语料库,以贝叶斯分类的消歧模式构造对维基百科的映射关系,并使用模式合并的方式形成新模式的方法来获取较高的准确率。宁博等[6]提出了基于异构知识库使用分布式计算的层次聚类方法,并在维基百科中文语料库的基础上融合了百科知识库。在 Hadoop 平台上用分布式计算进行层次聚类,研究人物实体特征的选取和维基百科等知识库的使用对命名实体消歧结果的影响。其中,通过加入百科知识库后,实验结果显示 F 值从原先的 91.33%提高到 92.68%。高艳红等[7]以中文维基百科为知识库支撑,从实体表述的语义环境和待选实体在百科中的描述两个方面提出不同的语义特征并计算语义相似度,在与已构建的图模型融合后,采取 PageRank 算法计算,F 值提高了 9%。马晓军等[8]针对 Skip-gram 词向量计算模型在处理多义词时只能计算一个混合多种语义的词向量,不能对多义词不同含义进行区分的问题,提出了与词向量相结合并融合了三类特征的方法,与此同时,择取待选实体中相似度最高的作为最终的目标实体取得了更优化的消歧结果。

目前,实体消歧技术虽然可以利用实体上下文和外部知识库来获取实体知识,但是面对短文本和缺乏知识库的情况,相关技术无法发挥其效果。为此,本文在没有外部知识库和丰富的上下文的情况下,提出一种新的按照领域划分的实体消歧技术。首先,为所有的待消歧词创建字典,以便快速精准地找到待消歧词;接着,对句子进行切分,每一个待消歧词切分为一句,防止因一句话中存在多个待消歧词的情况而出现判断误差;最后,将切分好的句子放入 BERT(bidirectional encoder representation from transformers)模型中进行预训练,对训练的结果进行分类,从而得出句子中是否存在领域实体。

本文结合 BERT 和 LSTM(long short-term memory)的优点,提出了一种基于深度学习的实体消歧技术,以提高消歧准确率。主要贡献如下:1)将二分类法引入实体消歧模型。实体消歧任务简化为以领域为单位,只判断待消歧实体是否为特定领域实体,使其相较于多分类法准确度大幅提升。2)将BERT模型和LSTM 模型相结合, 提取了具有较高准确度的特征向量。

1 背景知识

1.1 逻辑回归

在监督学习模型中,通过大量的训练数据,拟合出一个分类决策函数或者分类模型,并将此称作分类器。对于分类器而言,模型输入的变量既可以是连续的,也可以是离散的,而输出的变量则是有限个离散数值。对于训练好的分类器,当向模型输入新的变量时,模型可以输出预测结果,这一过程称作分类。

线性回归[9]是利用称为线性回归方程的最小二乘函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。

逻辑回归[9]是目前最成熟也是最为常见的二分类模型,主要用来计量一组解释变量与离散的因变量之间的关系。

对于二分类问题,结果只有 0,1 两种情形,可用 sigmoid 函数表示为

$$\mathrm{sigmoid}(x)=\frac{1}{1+\mathrm{e}^{-x}}。\tag{1}$$

对线性回归中的函数进行 sigmoid 变换, 即可得逻辑回归函数

$$h_\theta(x)=\mathrm{sigmoid}(\boldsymbol{\theta}^{\mathrm{T}}x)=\frac{1}{1+\mathrm{e}^{-\boldsymbol{\theta}^{\mathrm{T}}x}}。\tag{2}$$

南通大学学报(自然科学版)

对离散的数据进行sigmoid 变换后，可化为连续的线性回归，变换后可得

$$\ln\frac{y}{1-y}=\boldsymbol{\theta}^{\mathrm{T}}x。\tag{3}$$

根据概率，将等式(3)中的 y 替换为 $p(y=1|x)$，将 $1-y$ 替换为 $p(y=0|x)$，可解得

$$p(y=1|x)=\frac{e^{\boldsymbol{\theta}^{\mathrm{T}}x}}{1+e^{\boldsymbol{\theta}^{\mathrm{T}}x}},\tag{4}$$

$$p(y=0|x)=\frac{1}{1+e^{\boldsymbol{\theta}^{\mathrm{T}}x}}。\tag{5}$$

此时，只需求出 $\boldsymbol{\theta}$ 值，即可找出最优解，这里采用极大似然估计法对 $\boldsymbol{\theta}$ 进行确定，

$$L(\theta)=\sum_{i=1}^{n}[-y_i\boldsymbol{\theta}^{\mathrm{T}}x_i+\log(1+e^{\boldsymbol{\theta}^{\mathrm{T}}x_i})],\tag{6}$$

根据该函数可通过梯度下降法求出最优解。

1.2 BERT 模型

BERT[10]利用了 Transformer[11]的 encoder 部分，该模型的创新点主要在 Pre-train 上，即用 Masked LM 捕捉词语级别的 Representation 和用 NSP (next sentence prediction)方法捕捉句子级别的 Representation。Transformer 的原型包括两个独立的机制：Encoder 负责接收文本作为输入，Decoder 负责预测任务的结果。BERT 的目标是生成语言模型，所以只需要 Encoder 机制。

在将单词序列输入 BERT 之前，每个序列中有 15%的单词被[MASK] token 替换。然后模型尝试基于序列中其他未被 mask 的单词的上下文来预测被mask 的原单词。BERT 的损失函数只考虑了 mask 的预测值，而忽略了没有 mask 的字的预测。因此，双向模型要比单向模型收敛得慢，但结果的情境意识得到增强。

在 BERT 的训练过程中，模型接收成对的句子作为输入，并且预测其中第 2 个句子是否在原始文档中也是后续句子。在训练期间，50%的输入对在原始文档中是前后关系，另外 50%是从语料库中随机组成的，并且是与第一句断开的。

1.3 长短期记忆网络

在深度学习领域中，长短期记忆(long short-term memory，LSTM)[12]是一种循环神经网络(recurrent neural network，RNN)架构。相较于标准的前馈神经网络，LSTM 增加了反馈连接。因此，LSTM 不但可以处理类似图像的单个数据点的文件，同时对于语言抑或是视频这种整个数据序列文件也有显著的成效。LSTM 通常包括 1 个单元和 3 个门结构，其中：单元用来记录任何时间间隔内的相关值；3 个门分别为忘记门、输入门和输出门，它们控制信息流如何进出单元。由于存在于时间序列内的重要事件之间通常会出现未知的延续时间的滞后，LSTM 神经网络可以很好地对基于时间序列数据作出预测、分类和处理等操作。同时，相比于传统 RNN 在训练时出现梯度消失的问题，LSTM 可以通过门结构进而避免。

1.4 NNI

神经网络智能[13](neural network intelligence，NNI)是一个轻量级但功能强大的工具包，可帮助用户自动化功能工程、神经体系结构搜索、超参数调整和模型压缩。该工具可管理自动机器学习(AutoML)实验，调度并运行调整算法，以搜索不同训练环境中的最佳神经体系结构和/或超参数。

2 基于深度学习的实体消歧

鉴于BERT 模型具有并行、提取特征和对文本双向建模的能力，可以用较少的数据和较短的时间获得较好的结果，而长短期记忆神经网络可以保留较重要的信息、忘记冗余信息，本文将这两种技术结合起来并使用二分类技术对实体消歧，提出了一种新型的基于深度学习的实体消歧技术，其中包括两个阶段：特征工程和深度学习。该技术主要包括以下 4 个过程：1)通过jieba 分词技术，找出所有待消歧的实体；2)以待消歧词为中心，对句子进行切分；3)将切分好的句子放入已经预训练的BERT 模型，得到切分后的句子的词向量；4)将得到的词向量放入 LSTM 中，进行神经网络的训练，得到训练模型。

2.1 jieba 分词

本文为所有实体名创建了字典，再使用 jieba 分词技术[14]找出所有待消歧的实体。图 1 是 jieba 分词工作流程图。图中加载的字典为实体名，方便快速找出待消歧词。对待分词的文本生成前缀树，并用正则匹配构建潜在串序的有向无环图。通过动态规

南通大学学报（自然科学版）

划找出最大概率路径的分词方案,为了让分词效果适应文本,使用 Viterbi 算法求解 HMM(hidden Markov model)模型,挖掘新词。

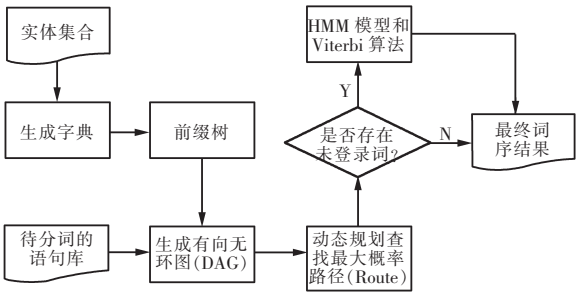

图 1 jieba 分词过程

Fig. 1 Word segmentation process of jieba

2.2 句子切分

由于长句子在神经网络训练时通常消耗较长的时间,所以本文对句子进行切分,且对句子进行编码时只选 32 个字,这样在保证准确率的基础上,尽可能地提高神经网络的训练速度。算法思想:以实体名为中心切分句子,先找到实体名在文本中的位置,再将实体名的前 13 个字和后 14 个字划分成一个句子,其中实体名固定占 5 个字节。算法 1 给出了句子切分算法。

算法 1 句子切分算法

输入:*train*

输出:*data2train*

FOR*i*IN*train*:

s = *str*(*i*[4]).*strip*() //原文本

comp = *str*(*i*[2]).*strip*() //文本中的实体名

npos = *s*.*find*(*comp*) //实体名在文本中的位置

pre = *npos* − 13

pos = *npos* + *len*(*comp*) + 14

IF*pre* < 0:

pos = *pos* − *pre*

pre = 0

pos = *min*(*len*(*s*), *pos*)

ELIF*pos* > *len*(*s*):

pre = *pre* − (*pos* − *len*(*s*))

pos = *len*(*s*)

pre = *max*(0, *pre*)

IF*npos* + *len*(*comp*) >= *len*(*s*):

new_s = *s*[*pre*:*npos*] + *comp*

ELSE:

new_s = *s*[*pre*:*npos*] + *comp* + *s*[*npos* + *len*(*comp*):*pos*]

data2train.*append*(*i*[0:4] + [*new_s*])

2.3 特征向量的获取

特征向量的获取分为两个步骤:预训练和LSTM神经网络训练。图 2 为获取过程的具体展示。首先,为了将文本信息转化为计算机可以识别的模式,模型分别将文本中的每个字转化为 BERT 模型对应该字的标识 ID,然后将每个标识 ID 映射到已经预训练好的 BERT 词向量。然后,将这些词向量分批次放入 LSTM 中进行神经网络训练,并以交叉熵作为损失函数,得到最终训练模型。

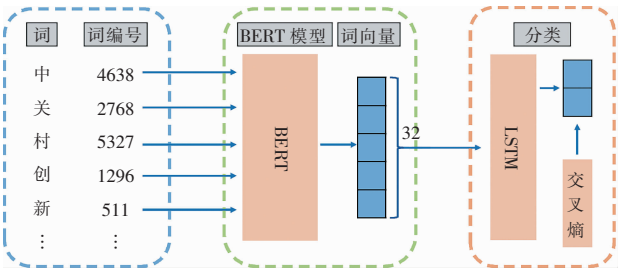

图 2 特征向量的获取

Fig. 2 Construction of features

2.4 句子的二分类

由于对于实体名来说,名称只是一个指示代词,并没有实际语法方面的意思,故本文将问题简化为二分类问题。这不仅简化了模型,提高了训练速度和准确率,也方便模型迁移到其他特殊领域的实体消歧问题。本文将是否为实体名进行二分类:是实体名为 1,非实体名为 0。

作为损失函数的一种,交叉熵是二分类的一种工具,它能衡量细微的差异、凸优化函数,便于利用梯度下降方法找到最优解。

交叉熵损失函数,其定义为

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[y_i \log(p_i) + (1-y_i)\log(1-p_i)], \quad (7)$$

式中:y_i 表示样本 i 的标签,正类为 1,负类为 0;p_i 表示样本 i 预测为正的概率。

南通大学学报(自然科学版)

3 实验分析

本实验主要研究如下3个问题：1)不同技术生成的词向量对最终实验准确率的影响；2)不同神经网络训练的最终模型效果比较；3)句子切分大小对神经网络的影响。

3.1 实验对象

本文通过网络爬虫技术，从金融领域的新闻页面中爬取长度不一的包含待消歧实体的文本作为本文的数据集；选择327个上市公司名称作为待消歧的对象。对于网络爬虫获得的数据进行人工标注，并将标注后的结果随机分为训练集和测试集，见表1。其中，正样本代表待消歧词是实体名，负样本代表待消歧词不是实体名。由于比例不平衡的样本数据会使得模型结果较差，而在实际应用中正样本数要比负样本数多，因此，本文在保障正负样本比例相对平衡的基础上，让正样本数略多于负样本数。

表1 实验对象

Tab. 1 Experiment subject

数据类型	正样本数	负样本数	总数
训练集	8 735	6 983	15 718
测试集	910	720	1 630

3.2 实验度量方法

本实验研究的3个问题最终评判的标准都是训练出的模型是否能够准确判断待消歧实体的属性(是否为实体)。由于在实际问题中，一个实体所表示的含义在不同场合下出现的频率是不同的，因此，只是单纯考虑准确率是不符合要求的，会使得小概率事件被忽略。本实验使用 F_1 值(F-Measure 值)作为评判标准，其定义如下：

$$P = \frac{n_{TP}}{n_{TP} + n_{FP}}, \tag{8}$$

$$R = \frac{n_{TP}}{n_{TP} + n_{FN}}, \tag{9}$$

$$F_1 = \frac{2PR}{P + R} = \frac{2n_{TP}}{2n_{TP} + n_{FP} + n_{FN}}。 \tag{10}$$

上述公式中，n_{TP} 表示正样本被判断为正的个数，n_{FP} 表示负样本被判断为正的个数，n_{FN} 表示正样本被判断为负的个数。公式(8)表示的是精准率，也就是被准确预测的正样本与所有被预测为正样本的个数之比。公式(9)表示的是召回率，也就是被准确预测的正样本与所有样本中事实是正样本的个数之比。公式(10)表示的是 F_1 值，F_1 值越大，模型表现效果越好。

3.3 实验过程

首先，利用Python的Scrapy包对金融领域新闻网页上的文本数据进行爬取，通过字符串匹配技术，提取本文所需要的含有待消歧词的文本，并通过人工标注的方法，判断待消歧词是否为实体；其次，将标注好标签的文本进行句子切分，缩小且统一句子长度；然后，将切分后的句子转化为词向量，放入神经网络中进行学习，获得训练后的模型；最后，将测试数据放入模型中获得 F_1 值，评判模型优劣。通过对Word2vec、BERT和ERNIE(enhanced language representation with informative entities)3种词向量技术和一般神经网络、卷积神经网络和长短期记忆神经网络3种神经网络架构 F_1 值的比较，验证本文所提方法的优越性。

3.4 实验结果与分析

3.4.1 实验结果

本实验中，BERT模型为谷歌预训练好的中文词向量，双向LSTM神经网络为两层，每层隐藏节点为768个，dropout为0.1，学习率为 5×10^{-6}。验证集的准确率和损失值如图3所示(每320个训练样本作为一个批次)。从图中可以看出，随着训练次数的增加，准确率不断提高而损失值不断下降且最终都趋于平稳。

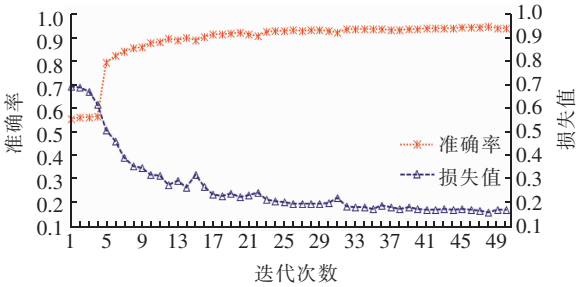

图3 验证集的准确率和损失值

Fig. 3 Accuracy and loss value of validation set

3.4.2 3种词向量模型的比较

对Word2vec、BERT和ERNIE 3种词向量模型进行比较，以测试集的 F_1 值作为判定依据。词向

南通大学学报（自然科学版）

量是通过训练将语言中的每一个词映射成一个固定长度的短向量,向量的每一个维度都有其特殊的含义,因此可以表达更多信息,同时,词向量还可以通过其空间距离来体现词与词之间的关系。对于Word2vec[15]模型来说,其关键思想是根据词的上下文语境来获得向量化表示,在本实验中,采取的是一种具有负采样的通过中心词预测附近词(Skip-gram)的方法。ERNIE 是对 BERT 模型的改进,它通过对训练数据中的词法结构、语法结构、语义信息统一建模,以此提高通用语义表示能力。实验过程中,Word2vec 使用的是 GitHub 开源的词向量,由于 Word2vec 最小单位是词而非字,故对处理后的文本使用 jieba 分词,使得每个词对应到相应的词向量。BERT 和 ERNIE 预训练模型分别来自 huggingface 和 nghuyong。BERT 模型和 ERNIE 模型的最小单位是字,故不需要进行分词处理。3 种词向量模型下得到的测试集 F_1 值如图 4 所示(每 320 个训练样本作为一个批次)。从图中可以看出,BERT 和 ERNIE 的结果最好,但 BERT 模型曲线更加平稳。

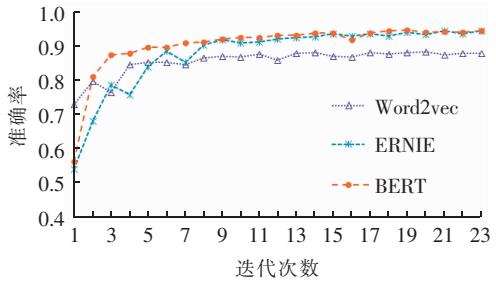

图 4 3 种词向量模型下得到的 F_1 值

Fig. 4 F_1 values of three word vectors

3.4.3 3 种神经网络模型的比较

对一般神经网络、卷积神经网络和长短期记忆神经网络 3 种模型[9,16]进行比较,以测试集的 F_1 值作为判定依据。作为一种运算训练模型,神经网络由大量的被称作节点的神经元相互连接组成。卷积神经网络作为一种前馈神经网络,它可以通过神经元来响应周围的神经单元,并且通常用于大型的图像处理任务。不同于 RNN 模型,LSTM 在其架构中增加了一个被称为 cell 的结构,它的作用是判断信息流是否有用。为了确保 3 种神经网络对于自身取得的都是最优结果,本文使用 NNI 工具进行调参。3 种神经网络模型下得到的测试集 F_1 值如图 5 所示(每 320 个训练样本作为一个批次)。从图中可以看出 LSTM 收敛更加平滑。

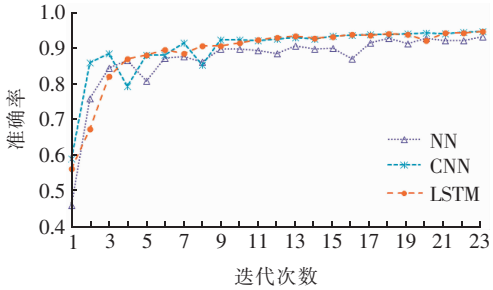

图 5 3 种神经网络模型下得到的 F_1 值

Fig. 5 F_1 values of three neural networks

3.4.4 不同文本长度的比较

对不同文本长度进行比较,以测试集的 F_1 值作为判定依据。虽然 LSTM 通过一种被称为门的结构对神经元状态进行删除或者添加信息,序列长度超过一定限度后,梯度还是会消失。然而,过短的序列会使得神经网络模型对文本中有用的信息无法充分获取,故在神经网络训练过程中文本的长度对其最终效果有着一定的影响。本小节主要对 3 种长度文本进行比较,其测试集比较结果如图 6 所示(每 320个训练样本作为一个批次)。由图可以看出,在相同的训练周期内,长度影响并不显著。

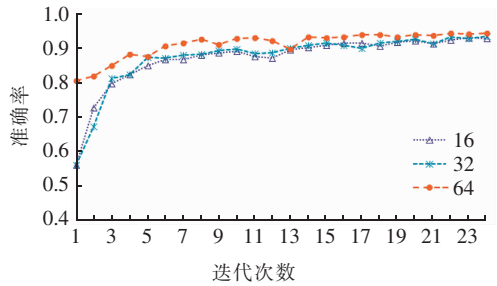

图 6 3 种文本长度下得到的 F_1 值

Fig. 6 F_1 values of three text lengths

3.5 讨论

通过上述实验结果与分析可知,使用 BERT 模型能够有效获取词之间的关系同时避免冗余信息的导入。对于神经网络,LSTM 的使用解决了长文本信息保存问题。此外,对文本长度的合理切分,可以获得足够多的信息同时训练速度得到提高。但在实验过程中仍然存在一些问题。

南通大学学报(自然科学版)

在合适的参数条件下，随着训练模型迭代次数的增加，不同的神经网络最终呈现的效果大体相似。然而，词向量的选择会对最终效果产生较大的影响。这说明文本研究的重点应该是如何为神经网络获得更多的预备知识，以此来更好地理解文本信息。

4 相关工作

随着自然语言处理领域不断的发展，实体消歧无疑是其中最为基础的研究对象。对于机器翻译系统，实体消歧通过特定的实体可以有效地选择最优翻译选项，优化翻译性能。对于知识图谱系统，精准的实体消歧可以保证实体间关联的正确性。对于推荐系统，通过对用户浏览信息文本进行分析并预测其中的大量待消歧词，系统才可以更好地获取用户兴趣[17]。实体消歧包含两种类型，分别是歧义性和多样性[7]。歧义性是指多个实体具有一样的命名，也就是一词多义；多样性是指一个实体具有多个命名，包括缩写、重名、别名等。

目前，实体消歧方法主要包括5种类型[18]，分别是实体显著性、上下文相似度、实体关联度、深度学习算法和特殊标识资源。基于实体显著性的命名实体识别是指从待选实体列表中选出显著性最高的作为结果。其中，显著性的含义有两种：一种是字符串相似度，另一种是流行度（或使用频率）。基于上下文相似度的命名实体消歧是指通过待选实体与实体所在的上下文文本进行相似度比较，并选择最好的结果。对于相似度的算法有两种：一种是词袋（bag-of-words，BOW）模型[19]，其中，使用词集合来表示实体所在的文本，相似度由词交集的大小来表示；另一种是向量空间模型，其权重主要以TF-IDF（term frequency-inverse document frequency）值来表示。为了提高结果准确率，目前也有许多模型被提出，如基于主题[20]、分类[21]、聚类[22-23]、概率语言模型[24-25]。不同于实体显著性和文本相似度的方法，实体关联度可以根据文本语义特征进行消歧，如使用协同消歧的方法[26]与使用隐马尔可夫模型的方法[27]，每次只能消歧一个实体。自从2013年Mikolov等[28]提出词向量以后，有关词向量软件工具被相继推出，如Word2vec、GloVe、BERT等。通过词向量和各种神经网络（LSTM、CNN、RNN）相结合构造出的模型对实体消歧结果有显著提高。特殊标识资源则是通过实体标识的语义来帮助进行实体消歧，这些标识通常是在某些领域通用的。

对于最近这些年关于实体消歧技术的发展，实体消歧依旧有许多问题需要解决。随着大数据相关技术的不断完善，深度学习对实体消歧有着举足轻重的作用，在特征提取、语义分析、模型优化等方面有着提升的空间。

5 结论

本文提出一种有效的基于深度学习的实体消歧技术。这种技术首先通过对长文本进行切分缩小为短文本，以减少神经网络的规模；其次，使用BERT模型作为词向量预训练模型，使得即使在较少的训练数据的情况下也可以获得较高的 F_1 值；最后，由于长文本的有用信息之间的距离较长，神经网络一般很难完全捕获，本文采用了长短期记忆神经网络技术，使用门结构来更好地保留信息。

本文的实验虽然在现实文本数据中验证了技术的有效性，但是基于深度学习的实体消歧技术像其他神经网络一样，对于与训练集属性差距较大的待消歧词，模型的最终 F_1 值会有所下降。下一步的工作将会着重研究如何更好地提取文本中的特征，以此来提高 F_1 值。

参考文献：

[1] 胡新辰. 基于LSTM的语义关系分类研究[D]. 哈尔滨：哈尔滨工业大学，2015.

[2] 张杨. 基于领域知识图谱实体消歧的协同过滤推荐算法研究[D]. 天津：天津师范大学，2019.

[3] 王博，杨沐昀，李生，等. 中文全词消歧在机器翻译系统中的性能评测[J]. 自动化学报，2008，34(5)：535-541.

[4] 左乃彻. 基于维基百科的中英文命名实体消歧[D]. 北京：北京邮电大学，2015.

[5] 邵发，黄银阁，周兰江，等. 基于实体消歧的中文实体关系抽取[J]. 山东大学学报(工学版)，2014，44(6)：32-37.

[6] 宁博，张菲菲. 基于异构知识库的命名实体消歧[J]. 西安邮电大学学报，2014，19(4)：70-76.

[7] 高艳红，李爱萍，段利国. 面向实体链接的多特征图模型实体消歧方法[J]. 计算机应用研究，2017，34(10)：2909-

2914.

[8] 马晓军, 郭剑毅, 王红斌, 等. 融合词向量和主题模型的领域实体消歧[J]. 模式识别与人工智能, 2017, 30(12): 1130-1137.

[9] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016:53-54.

[10] DEVLIN J, CHANG M, LEE K, et al. BERT:pre-training of deep bidirectional transformers for language understanding[DB/OL]. (2019-05-24)[2021-04-07]. https://arxiv.org/abs/1810.04805v2.

[11] 吴炎, 王儒敬. 基于 BERT 的语义匹配算法在问答系统中的应用[J]. 仪表技术, 2020(6):19-22.

[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[13] Microsoft. 概述:支持神经网络结构搜索、模型压缩、超参调优的开源自动机器学习工具(NNI v1.8)[EB/OL]. (2020-09-08)[2021-04-07]. https://nni.readthedocs.io/zh/latest/contents.html.

[14] Github. jieba[EB/OL]. (2020-02-15)[2021-04-07]. https://github.com/fxsjy/jieba/.

[15] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[DB/OL]. (2013-10-16)[2021-04-07]. https://arxiv.org/abs/1310.4546.

[16] VALUEVA M V, NAGORNOV N N, LYAKHOV P A, et al. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation[J]. Mathematics and Computers in Simulation, 2020, 177:232-243.

[17] 赵军. 命名实体识别、排歧和跨语言关联[J]. 中文信息学报, 2009, 23(2):3-17.

[18] 温萍梅, 叶志炜, 丁文健,等. 命名实体消歧研究进展综述[J]. 数据分析与知识发现, 2020, 4(9):15-25.

[19] HOFFART J, YOSE M A, BORDINO I, et al. Robust disambiguation of named entities in text[DB/OL]. (2011-07-27)[2021-04-07]. https://dl.acm.org/doi/abs/10.5555/2145432.2145521.

[20] 怀宝兴, 宝腾飞, 祝恒书, 等. 一种基于概率主题模型的命名实体链接方法[J]. 软件学报, 2014, 25(9):2076-2087.

[21] ZHANG W, SU J, TAN C L, et al. Entity linking leveraging automatically generated annotation[C]// The 23rd International Conference on Computational Linguistics Proceedings of the Main Conference (Volume 2). Beijing:Tsinghua University Press, 2010:619-627.

[22] 李广一, 王厚峰. 基于多步聚类的汉语命名实体识别和歧义消解[J]. 中文信息学报, 2013, 27(5):29-34.

[23] 谭咏梅, 杨雪. 结合实体链接与实体聚类的命名实体消歧[J]. 北京邮电大学学报, 2014, 37(5):36-40.

[24] HAN X P, SUN L. A generative entity-mention model for linking entities with knowledge base[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies. Stroudsburg, PA:Association for Computational Linguistics, 2011: 945-954.

[25] MEIJ E, BRON M, HOLLINK L, et al. Mapping queries to the Linking Open Data cloud:a case study using DBpedia[J]. Journal of Web Semantics, 2011, 9(4):418-433.

[26] CUCERZAN S. Large-scale named entity disambiguation based on wikipedia data[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague:Association for Computational Linguistics, 2007: 708-716.

[27] ALHELBAWY A, GAIZAUSKAS R. Named entity disambiguation using HMMs[C]//Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, November 17-20, 2013, Atlanta, GA, USA. Now York:IEEE Xplore, 2013: 159-162.

[28] MIKOLOV T, CHEN K , CORRADO G, et al. Efficient estimation of word representations in vector space[DB/OL]. (2013-09-07)[2021-04-07]. https://arxiv.org/abs/1301.3781v3.

(责任编辑:张燕)

南通大学学报(自然科学版)